Homogeneity Pursuit in Ranking Inferences Based on Pairwise Comparison Data

Yuxin Tao

Center for Statistical Science, Tsinghua University

(Joint work with **Prof. Zheng (Tracy) Ke, Harvard University**)

December 14, 2023

Applications of Ranking Inference

(a) Sports and Gaming Ranking
(b) Recommendation System and Web Search
(c) Journal Ranking, Univerisity Ranking, etc.

An example: NBA Basketball Team Ranking

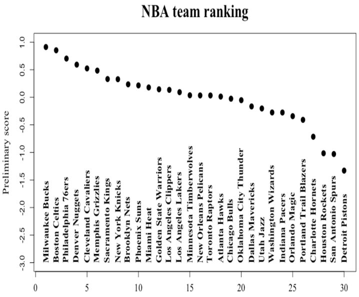


Fig 1. Estimated $\widehat{\theta}_i$ for 30 teams based on 2022-23 regular season.

- Data: Win/loss (binary data) in NBA games.
- n=30 teams, L=2-4 comparisons between each pair.
- (**Bradley-Terry-Luce**) Each team has a latent score θ_i.

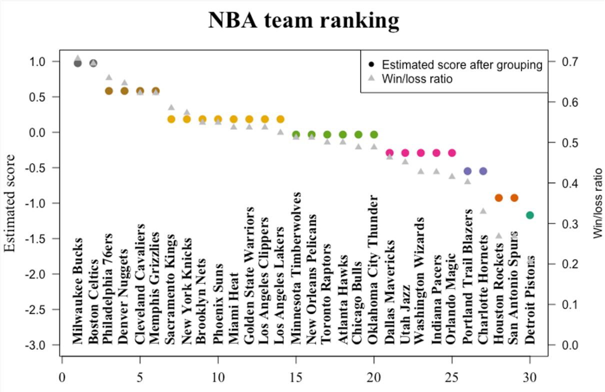
$$\mathbb{P}(i \text{ beats } j) = \frac{e^{\theta_i - \theta_j}}{1 + e^{\theta_i - \theta_j}}.$$

- Estimate $\boldsymbol{\theta}$ by MLE.
- Rank teams using $\widehat{\boldsymbol{\theta}}$.

Challenges and Open Questions

1. Potential overfitting due to insufficient comparisons.
   - It can be shown that $\|\widehat{\theta}_i - \theta_i\| \leq 1/\sqrt{nL}$. (Gao et.al., 2022)
   - For the NBA dataset, $n = 30$, $L \approx 3$, and $1/\sqrt{nL} \approx 0.1$. This is much larger than the difference between most adjacent $\widehat{\theta}_j$.
2. We are also interested in dividing teams into groups so that
   - There is no significant difference within each group.
   - There are significant differences between different groups.

For Today's Talk

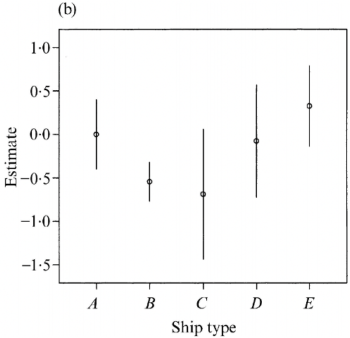


- **Model:** An extension of BTL with group structures.
- **Method:** Simultaneous parameter estimation and clustering.
- **Theory:** Faster rate than $1/\sqrt{nL}$.

Why our goal cannot be achieved from constructing confidence intervals?

- Many works on constructing confidence intervals for individual θ_i (Han et.al., 2020, Liu et.al., 2022, Gao et.al., 2022).
- An ad hoc approach: group i and j together if $\text{CI}_i \cap \text{CI}_j \neq \emptyset$.

The right shows the CIs in the application of ranking cargo ships' quality under wave damage incidents (Firth and Menezes, 2004). **All five CIs overlap**.



1. BTL Model with Group Structures

- n items
- Each item i is assigned with a latent preference score θ_i^*.
- $\mathbb{P}(i \text{ beats } j) \propto \exp(\theta_i^*)$, $\mathbb{P}(j \text{ beats } i) \propto \exp(\theta_j^*)$.
- $\mathbb{P}(i \text{ beats } j) = \frac{\exp(\theta_i^*)}{\exp(\theta_i^*)+\exp(\theta_j^*)} = \psi(\theta_i^* - \theta_j^*)$,
  where $\psi(\cdot)$ is the sigmoid function $\psi(t) \equiv e^t/(1+e^t)$.
  The log-odds is given by the difference of their scores.
- L independent comparisons for each observed pair (i, j):

$$y_{ij\ell} \overset{\text{ind}}{\sim} \text{Bernoulli}(\psi(\theta_i^* - \theta_j^*)), \quad l = 1, \ldots, L.$$

- Comparison graph $A_{ij} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$. Assume $p = 1$.

1. BTL Model with Group Structures

Suppose all individuals can be divided into K groups, a partition of $\{1, \ldots, n\}$, denoted as $G = (G_1, \cdots, G_K)$.
In each group, individuals share the same preference score,

$$\theta_j^* = \theta_{G,k}^*, \qquad \text{for all } j \in G_k, 1 \leq k \leq K.$$

Write $\boldsymbol{\theta}_G^* = (\theta_{G,1}^*, \ldots, \theta_{G,K}^*)^\top$ and $\boldsymbol{\theta}_G = (\theta_{G,1}, \ldots, \theta_{G,K})^\top$.

- WLOG, we assume $\theta_{G,1}^* < \theta_{G,2}^* < \cdots < \theta_{G,K}^*$.
- Each group has equal n/K individuals for notation simplicity in theoretical analysis.
- When $K = n$, it reduces to the standard BTL model.

1. BTL Model with Group Structures

Goal: conduct *estimation* and *inference* of the preference scores $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_n^*)^\top$ from pairwise comparisons.

1. BTL Model with Group Structures

Goal: conduct *estimation* and *inference* of the preference scores $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_n^*)^\top$ from pairwise comparisons.

Assumption 1

The parameter space for $\boldsymbol{\theta}^*$ is

$$\Theta(\kappa) = \left\{ \boldsymbol{\theta} \in \mathbb{R}^n : \max_{i \in [n]} \theta_i - \min_{i \in [n]} \theta_i \leq \kappa, \mathbf{1}_n^\top \theta = 0 \right\}.$$

- Here κ is known as the dynamic range, independent of n. We consider the fixed dynamic range regime, i.e., $\kappa = O(1)$.
- $\mathbf{1}_n^\top \boldsymbol{\theta}^* = 0$ for identifiability, as the BTL model is only identifiable up to a global shift in the score parameter $\boldsymbol{\theta}$.

2. Oracle Case

♣ When the group partition G is known,

Step 1: Compute $\bar{y}_{ij} = (\sum_{\ell=1}^{L} y_{ij\ell})/L$. (win ratio)

Step 2: Obtain the negative log-likelihood function

$$\begin{aligned} L_n(\boldsymbol{\theta}) &= \sum_{1\le i<j\le n}\left[\bar{y}_{ij}\log\frac{1}{\psi(\theta_i-\theta_j)}+\bar{y}_{ji}\log\frac{1}{\psi(\theta_j-\theta_i)}\right] \\ &= \sum_{1\le i<j\le K}\sum_{i'\in G_i,j'\in G_j}\left[\bar{y}_{i'j'}\log\frac{1}{\psi(\theta_{G,i}-\theta_{G,j})}+\bar{y}_{j'i'}\log\frac{1}{\psi(\theta_{G,j}-\theta_{G,i})}\right], \end{aligned}$$

where $\bar{y}_{ji} = 1-\bar{y}_{ij}$ by convention.

Step 3: Define the oracle MLE under the identifiability condition:

$$\widehat{\boldsymbol{\theta}}^{\text{oracle}} = \underset{\boldsymbol{\theta}:\mathbf{1}_n^\top\boldsymbol{\theta}=0}{\arg\min}\quad L_n(\boldsymbol{\theta}).$$

Proposition 1

Suppose the parameter space for $\boldsymbol{\theta}^*$ is $\Theta(\kappa)$, $\kappa = O(1)$. Assume the above assumptions hold and $K = o(n)$. Then we have

$$\|\widehat{\boldsymbol{\theta}}^{\text{oracle}} - \boldsymbol{\theta}^*\| \lesssim \sqrt{\frac{K + \log n}{Ln}}, \tag{1}$$

with probability at least $1 - O(n^{-7})$ uniformly over all $\boldsymbol{\theta}^* \in \Theta(\kappa)$.

Proposition (Existing results, e.g. Chen, Fan, Ma and Wang(2019), Chen, Gao and Zhang(2022))

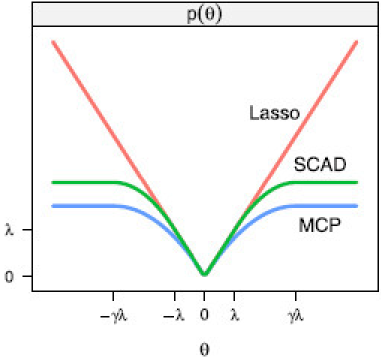
Assume $np \gtrsim log(n)$ (p=1 in our case), then w.h.p,

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \lesssim \sqrt{\frac{1}{L}}. \tag{2}$$

3. Methodology

♣ When the group partition G is unknown,

♣ Penalized MLE:
Likelihood function + Folded concave penalty (e.g., SCAD, Fan and Li 2001; MCP, Zhang 2010).
A symmetric function and nondecreasing and concave on $[0, \infty)$. There exists a constant $a > 0$ such that $\rho(\theta)$ is a constant for all $|\theta| \geq a\lambda$.



3. Methodology – Two possible penalties

♣ **The fused penalty with a pilot estimator**

Preordering: Construct the rank statistics $\{\tau(j) : 1 \leq j \leq n\}$ preliminary estimator $\boldsymbol{\theta}^{\mathsf{pre}}$, that is,

$$\theta^{\mathsf{pre}}_{\tau(1)} \leq \theta^{\mathsf{pre}}_{\tau(2)} \leq \cdots \leq \theta^{\mathsf{pre}}_{\tau(n)}.$$

$$\widehat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}:\mathbf{1}_n^T\boldsymbol{\theta}=0} \left\{ \frac{1}{n^2} L_n(\boldsymbol{\theta}) + \sum_{j=1}^{n-1} p_\lambda(|\theta_{\tau(j+1)} - \theta_{\tau(j)}|) \right\}.$$

♣ **The total variation penalty**

$$P_\lambda^{\mathrm{TV}}(\boldsymbol{\theta}) = \sum_{1\leq i,j \leq n} p_\lambda(|\theta_i - \theta_j|).$$

Methodology – Issues with fused penalty

Assumption 2

τ is consistent with the order of $\boldsymbol{\theta}^*$ with probability at least $1 - \epsilon_0$, that is,

$$\theta^*_{\tau(1)} \leq \theta^*_{\tau(2)} \leq \cdots \leq \theta^*_{\tau(n)}.$$

If the above assumption holds, then under some regularity conditions, $\widehat{\boldsymbol{\theta}}$ can consistently estimate the true coefficient groups of $\boldsymbol{\theta}^*$ with high probability.

Methodology – Issues with fused penalty

Assumption 2

τ is consistent with the order of $\boldsymbol{\theta}^*$ with probability at least $1 - \epsilon_0$, that is,

$$\theta^*_{\tau(1)} \leq \theta^*_{\tau(2)} \leq \cdots \leq \theta^*_{\tau(n)}.$$

If the above assumption holds, then under some regularity conditions, $\widehat{\boldsymbol{\theta}}$ can consistently estimate the true coefficient groups of $\boldsymbol{\theta}^*$ with high probability.

✘ Issues: Assumption 2 may be violated.

Methodology – Issues with fused penalty

Assumption 2

τ is consistent with the order of $\boldsymbol{\theta}^*$ with probability at least $1 - \epsilon_0$, that is,

$$\theta^*_{\tau(1)} \leq \theta^*_{\tau(2)} \leq \cdots \leq \theta^*_{\tau(n)}.$$

If the above assumption holds, then under some regularity conditions, $\widehat{\boldsymbol{\theta}}$ can consistently estimate the true coefficient groups of $\boldsymbol{\theta}^*$ with high probability.

✘ Issues: Assumption 2 may be violated.

✓ Solution: Use less information from $\boldsymbol{\theta}^{\text{pre}}$ and more penalty terms.

✓ Ke, Fan and Wu (2015): *Clustering Algorithm in Regression via Data-driven Segmentation (CARDS)*.

Ordered segmentation Υ

Let $\delta > 0$ be a pre-determined parameter, and find all indices $1 < i_2 < i_3 < \cdots < i_L$ such that the gaps

$$\theta^{\text{pre}}_{\tau(j)} - \theta^{\text{pre}}_{\tau(j-1)} > \delta, \quad j = i_2, \ldots, i_L.$$

Then, construct the segments

$$B_l = \{\tau(i_l), \tau(i_l + 1), \ldots, \tau(i_{l+1} - 1)\}, \quad l = 1, \ldots, L, \qquad (3)$$

where $i_1 = 1$ and $i_{L+1} = n + 1$.

Definition 1

Given a penalty function $p_\lambda(\cdot)$ and tuning parameters λ_1 and λ_2, the **hybrid pairwise penalty** corresponding to an ordered segmentation Υ is

$$\begin{aligned} P_{\Upsilon,\lambda_1,\lambda_2}(\boldsymbol{\theta}) = &\sum_{l=1}^{L-1} \sum_{i\in B_l, j\in B_{l+1}} p_{\lambda_1}(|\theta_i - \theta_j|) \\ &+ \sum_{l=1}^{L} \sum_{i,j\in B_l} p_{\lambda_2}(|\theta_i - \theta_j|). \end{aligned} \tag{4}$$

✓ Take advantage of the order of segments $B_1, \ldots, B_L$, and at the same time allow flexibility of order shuffling within each segment.

- When $L = n$, it reduces to the fused penalty.
- When $L = 1$, namely, no prior information about $\boldsymbol{\theta}$, (4) reduces to total variation penalty

How the Assumption 2 can be relaxed?

Definition 2

An ordered segmentation Υ **preserves the order** of $\boldsymbol{\theta}^*$ if $\max_{j \in B_l} \theta_j^* \le \min_{j \in B_{l+1}} \theta_j^*$, for $l = 1, \ldots, L-1$.

Assumption 3

The ordered segmentation Υ, generated by the preliminary estimator $\boldsymbol{\theta}^{\mathsf{pre}}$ and the tuning parameter δ_n, preserves the order of $\boldsymbol{\theta}^*$, with probability at least $1 - \epsilon_0$.

We group the coefficients $\boldsymbol{\theta}^{\mathsf{pre}}$ which differ by only a small amount into the same segment, therefore allowing some estimation error in preliminary ranking.

Methodology – CARDS

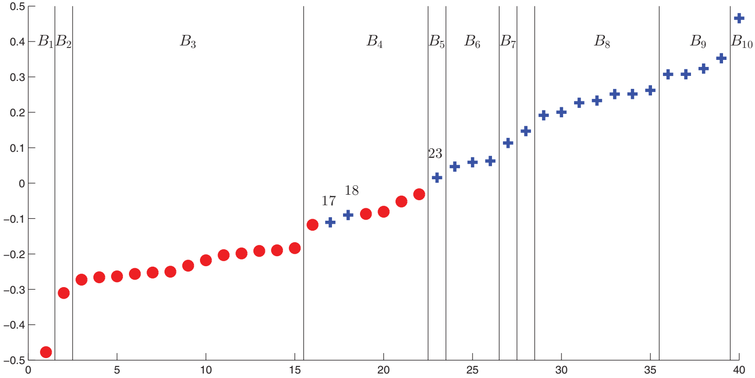


Figure 1: Illustration of CARDS penalty.

Methodology – CARDS

Procedure of CARDS:

- Preliminary Ranking: Given a preliminary estimate $\boldsymbol{\theta}^{\text{pre}}$, generate the rank mapping $\{\tau(j) : 1 \le j \le n\}$ such that $\theta^{\text{pre}}_{\tau(1)} \le \theta^{\text{pre}}_{\tau(2)} \le \cdots \le \theta^{\text{pre}}_{\tau(n)}$.
- Segmentation: For a tuning parameter $\delta > 0$, construct an ordered segmentation Υ as described in (3).
- Estimation: For tuning parameters λ_1 and λ_2, compute the solution $\widehat{\boldsymbol{\theta}}$ that minimizes

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n^2} L_n(\boldsymbol{\theta}) + P_{\Upsilon,\lambda_1,\lambda_2}(\boldsymbol{\theta}), \tag{5}$$

where $$L_n(\boldsymbol{\theta}) = \sum_{1 \le i < j \le n} \left[\bar{y}_{ij} \log \frac{1}{\psi(\theta_i - \theta_j)} + \bar{y}_{ji} \log \frac{1}{\psi(\theta_j - \theta_i)} \right].$$

Remark: Fused penalty is a special case of CARDS with $\delta = 0$.

4. Theory – Properties of CARDS

For given $G_1, \ldots, G_K$ and a segmentation $\Upsilon = \{B_1, \ldots, B_L\}$, define

$$\phi_k = |G_k| / \min\left\{ |G_k|^2, \min_{l: B_l \cap G_k \neq \emptyset} \{|B_l|^2\} \right\}.$$

Here $1/|G_k| \leq \phi_k \leq |G_k|$ for $1 \leq k \leq K$.

Assumption (4)

The ordered segmentation Υ, generated by the preliminary estimator $\boldsymbol{\theta}^{pre}$ and the tuning parameter δ_n, preserves the order of $\boldsymbol{\theta}^*$, with probability at least $1 - \epsilon_0$.

Properties of CARDS

Theorem 1

Suppose the above assumptions hold, $K = o(n)$. If the half minimum gap between groups, b_n, satisfies that $b_n > a \max\{\lambda_{1n}, \lambda_{2n}\}$, and

$$\lambda_{1n} \gg \max_k \Big\{ \frac{C\phi_k}{n} \sqrt{\frac{\log n}{Ln}} + C\sqrt{\frac{K + \log n}{Ln}} \Big\}, \qquad (6)$$

$$\lambda_{2n} \gg \max_k \Big\{ \frac{C}{n|G_k|} \sqrt{\frac{\log n}{Ln}} + C\sqrt{\frac{K + \log n}{Ln}} \Big\}, \qquad (7)$$

then with probability at least $1 - \epsilon_0 - cn^{-7}$, the CARDS objective function (5) has a strictly local minimizer $\widehat{\boldsymbol{\theta}}$ such that

- $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}^{\text{oracle}}$,
- $\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\| = O_p(\sqrt{(K + \log n)/(Ln)})$.

Properties of fused penalty

Theorem 2

Suppose the above assumptions hold, $K = o(n)$. If the half minimum gap between groups, b_n, satisfies that $b_n > a\lambda_n$, and

$$\lambda_n \gg C\frac{\max_k |G_k|}{n}\sqrt{\frac{\log n}{Ln}} + C\sqrt{\frac{K + \log n}{Ln}}, \tag{8}$$

then with probability at least $1 - \epsilon_0 - cn^{-7}$, the objective function with fused penalty has a strictly local minimizer $\widehat{\boldsymbol{\theta}}$ such that

- $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}^{\text{oracle}}$,
- $\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\| = O_p(\sqrt{(K + \log n)/(Ln)})$.

5. Real Data – NBA Basketball Team Ranking

- n=30 teams, L=2-4, total games=1230.
- Each pair of teams play at least 2 games, and at most 4 games.
- Each team plays 82 games, 41 each home and away.

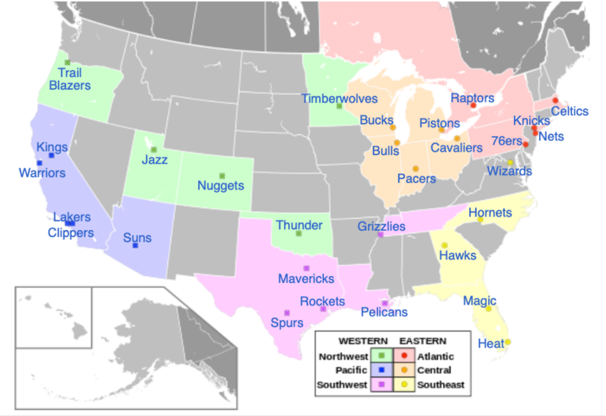


Figure 2: NBA basketball team.

NBA Basketball Team Ranking

♣ 2022-23 NBA regular season results

Conference Standings

* Playoff teams

Eastern Conference	W	L	W/L%	GB	PS/G	PA/G	SRS
Milwaukee Bucks*	58	24	.707	—	116.9	113.3	3.61
Boston Celtics*	57	25	.695	1.0	117.9	111.4	6.38
Philadelphia 76ers*	54	28	.659	4.0	115.2	110.9	4.37
Cleveland Cavaliers*	51	31	.622	7.0	112.3	106.9	5.23
New York Knicks*	47	35	.573	11.0	116.0	113.1	2.99
Brooklyn Nets*	45	37	.549	13.0	113.4	112.5	1.03
Miami Heat*	44	38	.537	14.0	109.5	109.8	-0.13
Atlanta Hawks*	41	41	.500	17.0	118.4	118.1	0.32
Toronto Raptors*	41	41	.500	17.0	112.9	111.4	1.59
Chicago Bulls*	40	42	.488	18.0	113.1	111.8	1.37
Indiana Pacers	35	47	.427	23.0	116.3	119.5	-2.91
Washington Wizards	35	47	.427	23.0	113.2	114.4	-1.06
Orlando Magic	34	48	.415	24.0	111.4	114.0	-2.39
Charlotte Hornets	27	55	.329	31.0	111.0	117.2	-5.89
Detroit Pistons	17	65	.207	41.0	110.3	118.5	-7.73

Western Conference	W	L	W/L%	GB	PS/G	PA/G	SRS
Denver Nuggets*	53	29	.646	—	115.8	112.5	3.04
Memphis Grizzlies*	51	31	.622	2.0	116.9	113.0	3.60
Sacramento Kings*	48	34	.585	5.0	120.7	118.1	2.30
Phoenix Suns*	45	37	.549	8.0	113.6	111.6	2.08
Los Angeles Clippers*	44	38	.537	9.0	113.6	113.1	0.31
Golden State Warriors*	44	38	.537	9.0	118.9	117.1	1.66
Los Angeles Lakers*	43	39	.524	10.0	117.2	116.6	0.43
Minnesota Timberwolves*	42	40	.512	11.0	115.8	115.8	-0.22
New Orleans Pelicans*	42	40	.512	11.0	114.4	112.5	1.63
Oklahoma City Thunder*	40	42	.488	13.0	117.5	116.4	0.96
Dallas Mavericks	38	44	.463	15.0	114.2	114.1	-0.14
Utah Jazz	37	45	.451	16.0	117.1	118.0	-1.03
Portland Trail Blazers	33	49	.402	20.0	113.4	117.4	-3.96
Houston Rockets	22	60	.268	31.0	110.7	118.6	-7.62
San Antonio Spurs	22	60	.268	31.0	113.0	123.1	-9.82

NBA Basketball Team Ranking

Preliminary team ranking: (teams with larger scores θ rank first)

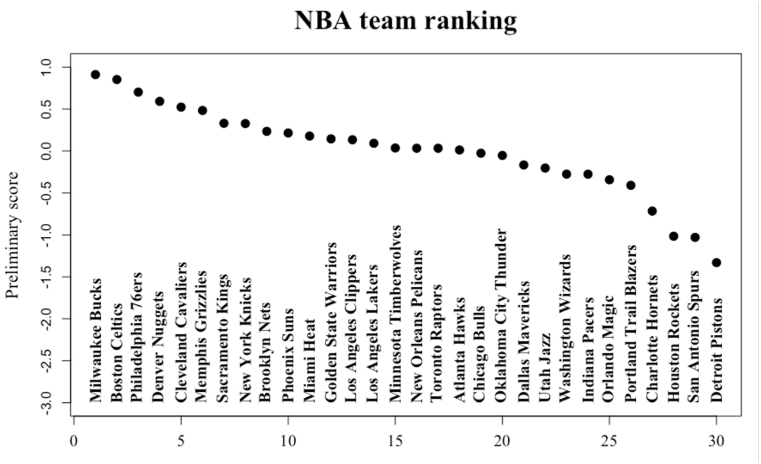
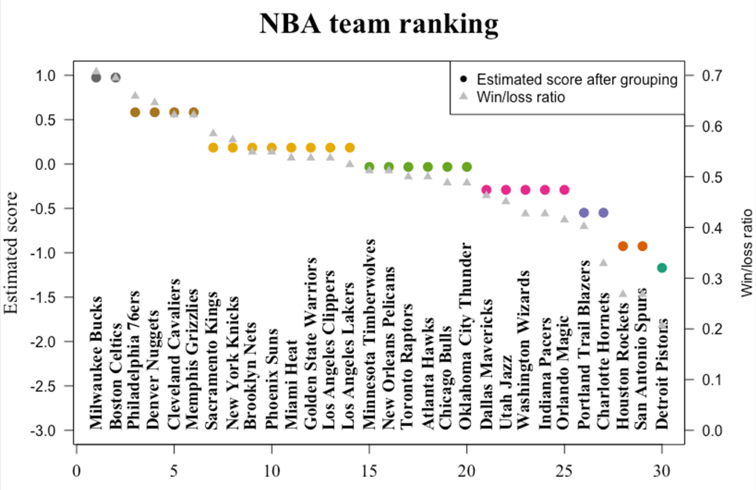


Figure 3: Preliminary estimation for NBA basketball team.

NBA Basketball Team Ranking

Team ranking using CARDS under homogeneity assumption:

- Use SCAD penalty, a=3.7. Group number $K = 8$.
- Ranking is consistent with the win/loss ratio for each team.



NBA Basketball Team Ranking

- Prediction error = average of square of $(y - \hat{y})$, where $\hat{y}$ is the estimated probability of wining.
- 40 random splits, training data: 60% (80%); testing data: 40% (20%).
- Comparison with pure BTL with no penalty:

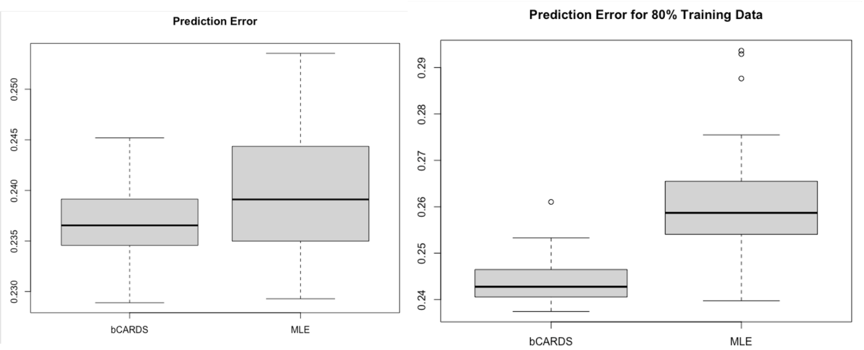


Figure 5: Prediction with 60% and 80% training data.

5. Real Data – Journal Ranking

- MADStat.
- Papers in one journal tend to cite those papers from journal with a higher prestige.
- n=33 (exclude three probability journals AIHPP, AoP, PTRF)
- Total citations between different journals=25248, citations using 10-year window, summation of year 2014 and 2015.

Journal Ranking

Preliminary movie ranking: (journals with smaller scores θ rank first)

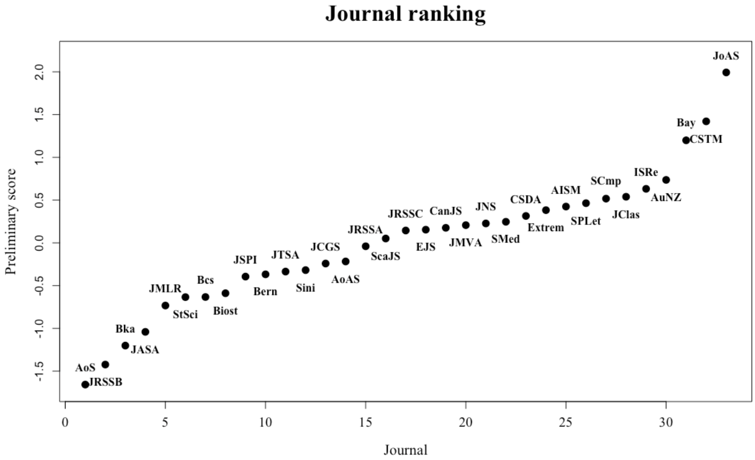
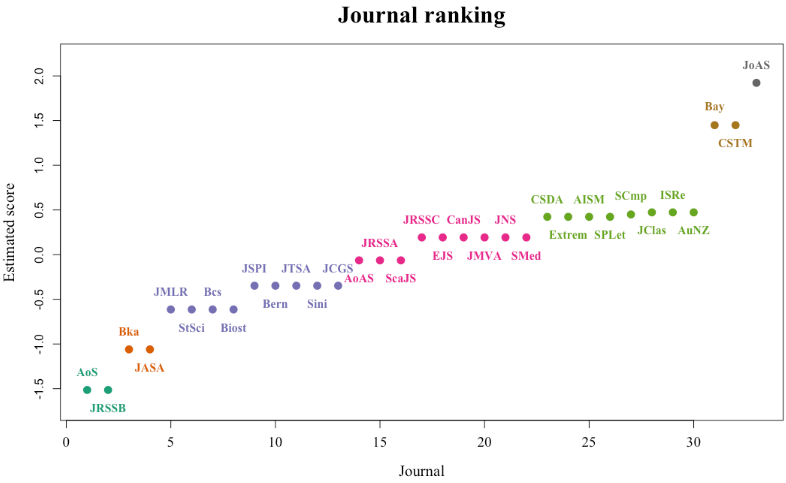


Figure 6: Preliminary estimation.

Netflix Film Ranking

Team ranking using CARDS under homogeneity assumption:

- Choose λ based on cross-validation error.
- Group number $K = 11$.



5. Real Data – Netflix Film Ranking

- Dataset: "Netflix prize dataset". Netflix held the Netflix Prize open competition for the best algorithm to predict user ratings for films. A total of 17770 movies from 1915-2005, and a total of 143458 reviewers.
- Extract n=100 movies, with highest number of reviews.
- Extract 52064 users rating info, who have rated more than 50 movies among these 100 movies.
- For each pair of movies, randomly select L=50 ratings for comparison.

Netflix Film Ranking

Preliminary movie ranking: (movies with larger scores θ rank first)

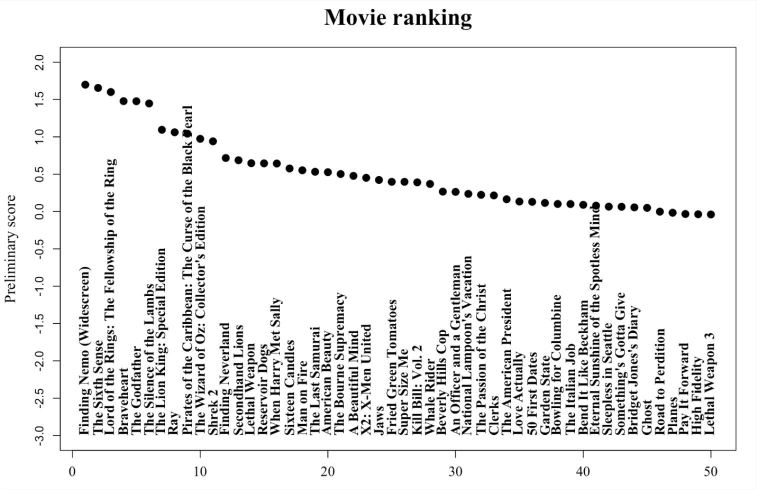


Figure 7: Preliminary estimation for Netflix film.

Netflix Film Ranking

Team ranking using bCARDS under homogeneity assumption:

- Choose λ based on cross-validation error.
- Use SCAD penalty, a=3.7. Group number $K = 19$.

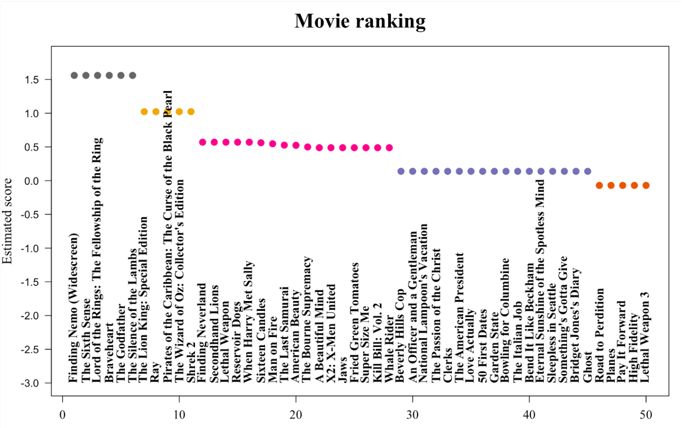


Figure 8: Movie ranking using CARDS under homogeneity assumption.

Netflix Film Ranking

Top six Netflix films:

- Finding Nemo
- The Sixth Sense
- Lord of the Rings: The Fellowship of the Ring
- Braveheart
- The Godfather
- The Silence of the Lambs

5. Summary

- We explore the homogeneity of scores in the BTL model, which assume that individuals cluster into group with the same preference scores.
- Introduce CARDS penalty to estimate scores and group structures at the same time.
  - More rigorous in methodology.
  - Obtain faster convergence rate and sharper confidence intervals.
  - Improve the prediction performance.
  - Allow bias in the order of preliminary estimation.
- Statistical properties of CARDS.
- Real data analyses including sports and movies ranking to demonstrate the efficiency and interpretation ability of our model.
- (Ongoing) Ranking inference – sharper confidence intervals.

Thank you!

Ke, Z.T. and **Tao, Y.**† (2023). Homogeneity pursuit in ranking inferences based on pairwise comparison data. *Manuscript.*

References

- Chen, Y., Fan, J., Ma, C., and Wang, K. (2019). Spectral method and regularized MLE are both optimal for top-K ranking. *The Annals of Statistics*, **47**, 2204–2235.
- Chen, P., Gao, C., and Zhang, A. Y. (2022). Partial recovery for top-k ranking: optimality of MLE and suboptimality of the spectral method. *The Annals of Statistics*, **50**(3), 1618–1652.
- Fan, J., and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of American Statistical Association*, **96**, 1348–1360.
- Gao, C., Shen, Y., and Zhang, A.Y. (2023). Uncertainty quantification in the Bradley-Terry-Luce model. *Information and Inference: A Journal of the IMA*, **12**, 1073–1140.
- Harchaoui, Z., and Lévy-Leduc, C. (2010). Multiple Change-Point Estimation With a Total Variation Penalty. *Journal of the American Statistical Association*, **105**, 1480–1493.
- Ke, Z.T., Fan, J., and Wu, Y. (2015). Homogeneity Pursuit. *Journal of the American Statistical Association*. **110**:509, 175–194.